Feature Selection for Aiding Glass Forensic Evidence Analysis

Shen, Qiang; Jensen, Richard

Published in:
Intelligent Data Analysis

DOI:
10.3233/IDA-2009-0389

Publication date:
2009

Citation for published version (APA):
Shen, Q., & Jensen, R. (2009). Feature Selection for Aiding Glass Forensic Evidence Analysis. *Intelligent Data Analysis*, *13*(5), 703-723. https://doi.org/10.3233/IDA-2009-0389

Feature Selection for Aiding Glass Forensic Evidence Analysis

Richard Jensen*and Qiang Shen
{rkj, qqs}@aber.ac.uk
Department of Computer Science, Aberystwyth University, Wales, UK

Abstract

The evaluation of glass evidence in forensic science is an important issue. Traditionally, this has depended on the comparison of the physical and chemical attributes of an unknown fragment with a control fragment. A high degree of discrimination between glass fragments is now achievable due to advances in analytical capabilities. A random effects model using two levels of hierarchical nesting is applied to the calculation of a likelihood ratio (LR) as a solution to the problem of comparison between two sets of replicated continuous observations where it is unknown whether the sets of measurements shared a common origin. Replicate measurements from a population of such measurements allow the calculation of both within-group and between-group variances. Univariate normal kernel estimation procedures have been used for this, where the between-group distribution is considered to be non-normal. However, the choice of variable for use in LR estimation is critical to the quality of LR produced. This paper investigates the use of feature selection for the purpose of selecting the variable for estimation without the need for expert knowledge. Results are recorded for several selectors using normal, exponential, adaptive and biweight kernel estimation techniques. Misclassification rates for the LR estimators are used to measure performance. The experiments performed reveal the capability of the proposed approach for this task.

Keywords

Feature selection; Fuzzy-rough sets; Glass analysis; Forensic evidence; Two-level model

*corresponding author

1 Introduction

One of the less obvious, but frequent, sources of forensic evidence are traces of glass. This is regularly encountered at crime scenes, particularly those involving motor vehicle accidents, car theft and burglaries. Windows are a common point of entry into buildings for burglars and large quantities of broken glass are produced in traffic accidents. Such glass fragments may remain for a long time (depending on the type of crime) and do not degrade like biological evidence. In addition fragments may also be transferred to anyone present during the glass breakage, or even to someone having secondary contact with the offender.

The forensic scientist's role in analysing glass is to clearly and unambiguously determine the origin of the sample. Variation in the manufacture of glass allows considerable discrimination even with very small fragments. Consider the scenario where a crime has been committed involving the breakage of glass. It is likely that fragments will remain at the location of the offence (referred to as control fragments). Fragments may also have been transferred to the clothes and footwear of the offender. A suspect may be identified and, on subsequent examination, found to have glass fragments on their person (referred to as recovered fragments as their source is not known). In this case, the purpose of fragment analysis is to evaluate the evidence for comparison of the proposition that the glass associated with the suspect is from the same source as the fragments from the crime scene with the proposition that the glass associated with the suspect is not from the same source as the fragments from the crime scene.

The first stage of the examination of this kind of evidence is the recovery of

glass fragments from the suspect. This is most frequently achieved through shaking and/or brushing garments. The resulting debris is observed under an optical microscope and glass fragments separated manually. The physico-chemical properties of the resulting fragments are then determined, often through the GRIM (Glass Refractive Index Measurements) method and instrumental methods of elemental assay (e.g. μ-XRF, LA-ICP-MS, SEM-EDX). The comparison between recovered and control glass fragments is then made on the basis of the analytical results.

The increasing ability to collect and store data relevant for identification in a forensic context has led to a corresponding increase in methods for the numerical evaluation of evidence associated with particular evidence types. The comparison of two sets of glass fragments by numerical methods requires careful attention to the following considerations. Firstly, the similarity of recovered glass fragment(s) to a control sample must be taken into account. Secondly, information must be considered about the rarity of the determined physico-chemical characteristics (e.g. elemental concentrations) for control and recovered samples in the relevant population. Thirdly, the level of association between different characteristics where more than one characteristic has been measured should be accounted for. Fourthly, other possible sources of variation should be considered, including the variation of measurements of characteristics within the control items, within recovered items, and between control and recovered items.

Significance tests are often used for this purpose. However, these only take into account information concerning within-source variation and item similar-

ity. This is a comparison of the samples based purely on their physico-chemical properties. From a forensic viewpoint, knowledge concerning the sources of variability and rarity of the measured properties should also be considered to produce a more definitive evaluation. This motivates the use of two-level univariate models that can incorporate such information effectively [1].

As discussed in [2], problems are encountered when dealing with multivariate forensic data. When modelling multivariate databases, there is a lack of background data from which to estimate the parameters of the assumed distributions such as means, variances and covariances. For example, when glass samples are described by seven variables then it is necessary to estimate, reliably, seven means, seven variances and 21 covariances for both within-group objects, and between-group objects. This requires far more analytical data than is accessible in many forensic databases, and observation of more variables, which in applied forensic contexts may be required, would necessitate the estimation of an exponentially larger number of means, variances, and covariances. The calculation of a full model is simply not practical.

As a result of these issues, a single informative variable is usually selected by an expert for use in the univariate model. This is the method followed by the work in [3], where modeling takes place once one feature has been chosen by the domain expert. The choice of such a variable is obviously a critical factor in the resulting quality of evidence evaluation. Unfortunately, it is not always known which single feature will provide the most information for this purpose. There are also situations where many competing variables co-exist;

manual selection of which variable to use may result in subsequent analysis being too subjective. Through the use of feature selection methods, this important decision can be made without expert knowledge. Indeed, experts may not be available for making such decisions. The opposing side in a court case could argue that, by selecting features, there is a loss of information. However, it is often the case that datasets contain many features that are purely redundant. The inclusion of such features will degrade the resulting quality of analysis. It is useful in this case to remove these or at least recommend these for potential removal to the forensic scientist.

This paper investigates and compares the effectiveness of different feature selection methods and likelihood ratio estimation procedures for the glass forensic evidence analysis domain. The rest of this paper is structured as follows. The second section describes the recently developed fuzzy-rough set-based feature selection metric in addition to the current leading measures in the field. Section three outlines the new two-level univariate estimation techniques. Section four describes the application of the present work to glass analysis, and the fifth section details the experimentation carried out and shows the results of applying the techniques to this domain. The final section concludes the paper, and proposes further work in this area.

2 Feature selection

The main aim of feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the

original features. In many real world problems feature selection is a must due to the abundance of noisy, irrelevant or misleading features. For instance, by removing these factors, learning from data techniques can benefit greatly. A detailed review of feature selection techniques devised for classification tasks and statistical methods can be found in [10, 17, 25].

2.1 Fuzzy-rough feature selection

Rough set theory (RST) [4, 27] can be used as such a tool to discover data dependencies and to reduce the number of attributes contained in a dataset using the data alone, requiring no additional information [15]. Over recent years, RST has indeed become a topic of great interest to researchers and has been applied to many domains. Given a dataset with discretized attribute values, it is possible to find a subset (termed a *reduct*) of the original attributes using RST that are the most informative; all other attributes can be removed from the dataset with very little information loss.

The rough set-based selection process described in [5] can only operate effectively with datasets containing discrete values. Additionally, there is no way of handling noisy data. As most datasets contain real-valued attributes, it is necessary to perform a discretization step beforehand [24]. This can be implemented by standard fuzzification techniques [30], enabling linguistic labels to be associated with attribute values. It also aids the modelling of uncertainty in data by allowing the possibility of the membership of a value to more than one fuzzy label. However, membership degrees of attribute values to fuzzy sets are not exploited in the process of dimensionality reduction. By using *fuzzy-*

rough sets [8], it is possible to use this information to better guide attribute selection. Here, real-valued attributes are still used, but are augmented with corresponding fuzzy set definitions.

2.1.1 Fuzzy equivalence classes

In crisp rough set theory, equivalence classes (obtained through an indiscernibility relation) are used to approximate concepts. Two objects are indiscernible (according to some subset of features) if their values are identical for this subset. In the same way that these equivalence classes are central to crisp rough sets [27], *fuzzy* equivalence classes are central to the fuzzy-rough set approach [8]. For typical applications, this means that the decision values and the conditional values may all be fuzzy. The family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes [8].

2.1.2 Fuzzy lower and upper approximations

The fuzzy lower and upper approximations are fuzzy extensions of their crisp counterparts. Informally, in crisp rough set theory, the lower approximation of a set contains those objects that belong to it with certainty. The upper approximation of a set contains the objects that possibly belong. The definitions given in [8] diverge a little from the crisp upper and lower approximations, as the memberships of individual objects to the approximations are not explicitly available. As a result of this, the fuzzy lower and upper approximations are redefined as:

$$\mu_{\underline{P}X}(x) = \sup_{F \in \mathbb{U}/P} \min(\mu_F(x), \inf_{y \in \mathbb{U}} \max\{1 - \mu_F(y), \mu_X(y)\}) \quad (1)$$

$$\mu_{\overline{P}X}(x) = \sup_{F \in \mathbb{U}/P} \min(\mu_F(x), \sup_{y \in \mathbb{U}} \min\{\mu_F(y), \mu_X(y)\}) \quad (2)$$

where $\mu_F(x)$ is the degree of membership of x to fuzzy equivalence class F, and $\mu_X(x)$ is the degree of membership of x to the decision concept X. The tuple $\langle \underline{P}X, \overline{P}X \rangle$ is called a fuzzy-rough set.

For an individual feature, a, the partition of the universe by $\{a\}$ (denoted $\mathbb{U}/IND(\{a\})$) is considered to be the set of those fuzzy equivalence classes for that feature. For subsets of feature, the following is used:

$$\mathbb{U}/P = \otimes\{a \in P : \mathbb{U}/IND(\{a\})\} \quad (3)$$

Each set in $\mathbb{U}/P$ denotes an equivalence class. The extent to which an object belongs to such an equivalence class is therefore calculated by using the conjunction of constituent fuzzy equivalence classes, say F_i, $i = 1, 2, ..., n$:

$$\mu_{F_1 \cap ... \cap F_n}(x) = \min(\mu_{F_1}(x), \mu_{F_2}(x), ..., \mu_{F_n}(x)) \quad (4)$$

2.1.3 Fuzzy-rough reduction process

Fuzzy-Rough Feature selection (FRFS) [15] builds on the notion of the fuzzy lower approximation to enable reduction of datasets containing real-valued features. The process becomes identical to the crisp approach when dealing with nominal well-defined features.

The crisp positive region in the standard RST is defined as the union of the lower approximations. By the extension principle, the membership of an object $x \in \mathbb{U}$, belonging to the fuzzy positive region can be defined by

$$\mu_{POS_P(Q)}(x) = \sup_{X \in \mathbb{U}/Q} \mu_{\underline{P}X}(x) \tag{5}$$

Using the definition of the fuzzy positive region, a new dependency function between a set of features Q and another set P can be defined as follows:

$$\gamma'_P(Q) = \frac{|\mu_{POS_P(Q)}(x)|}{|\mathbb{U}|} = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_P(Q)}(x)}{|\mathbb{U}|} \tag{6}$$

As with crisp rough sets, the dependency of Q on P is the proportion of objects that are discernible out of the entire dataset. In the present approach, this corresponds to determining the fuzzy cardinality of $\mu_{POS_P(Q)}(x)$ divided by the total number of objects in the universe. This function can be used to evaluate individual features to generate rankings, or can be used to guide subset search as part of a feature selection process. When requiring feature selection rather than individual feature evaluation, a greedy hill-climbing algorithm is typically used [5].

2.2 Fuzzy entropy measure

Again, let $I = (\mathbb{U}, \mathbb{A})$ be a decision system, where $\mathbb{U}$ is a non-empty set of finite objects. $\mathbb{A} = \{\mathbb{C} \cup \mathbb{D}\}$ is a non-empty finite set of attributes, where $\mathbb{C}$ is the set of input features and $\mathbb{D}$ is the set of decision features. An attribute $a \in \mathbb{A}$ has corresponding fuzzy subsets $F_1, F_2, ..., F_n$. The fuzzy entropy for a fuzzy subset

F_i can be defined as:

$$H(\mathbb{D}|F_i) = \sum_{Q \in \mathbb{U}/\mathbb{D}} -p(Q|F_i) \log_2 p(Q|F_i) \tag{7}$$

where, $p(Q|F_i)$ is the relative frequency of the fuzzy subset F_i of attribute a with respect to the decision Q, and is defined:

$$p(Q|F_i) = \frac{|Q \cap F_i|}{|F_i|} \tag{8}$$

The cardinality of a fuzzy set is denoted by $|\cdot|$. Based on these definitions, the fuzzy entropy for an attribute subset R is defined as follows:

$$E(\mathbb{D}|R) = \sum_{F_i \in \mathbb{U}/R} \frac{|F_i|}{\sum_{Y_i \in \mathbb{U}/R} |Y_i|} H(\mathbb{D}|F_i) \tag{9}$$

This fuzzy entropy can be used to gauge the utility of attribute subsets in a similar way to that of the fuzzy-rough measure. However, the fuzzy entropy measure decreases with increasing subset utility, whereas the fuzzy-rough dependency measure increases.

2.3 Further metrics

Other leading feature significance measures in this field are presented here, for use in the application. Information gain, gain ratio, χ^2 and symmetrical uncertainty are used to evaluate individual features. Relief-F and OneR incorporate feature evaluation as part of their overall feature selection process. Further details concerning the operation of these methods can be found in the associated references.

2.3.1 Information gain

The Information Gain (IG) [13] is the expected reduction in (crisp) entropy resulting from partitioning the dataset objects according to a particular feature. The entropy of a labelled collection of objects S is defined as:

$$Entropy(S) = \sum_{i=1}^{c} -p_i log_2 p_i \tag{10}$$

where p_i is the proportion of S belonging to class i. Based on this, the Information Gain metric is:

$$IG(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{11}$$

where $values(A)$ is the set of values for feature A, S the set of training examples, S_v the set of training objects where A has the value v. This metric is the one used in ID3 [29] for selecting the best feature to partition the data.

2.3.2 Gain ratio

One limitation of the IG measure is that it favours features with many values. The Gain Ratio (GR) seeks to avoid this bias by incorporating another term, split information, that is sensitive to how broadly and uniformly the attribute splits the considered data:

$$Split(S, A) = - \sum_{v \in values(A)} \frac{|S_v|}{|S|} log_2 \frac{|S_v|}{|S|} \tag{12}$$

where each S_v is the subset of training objects where A has value v. The Gain Ratio is then defined as follows:

$$GR(S, A) = \frac{IG(S, A)}{Split(S, A)} \tag{13}$$

2.3.3 χ^2 measure

In the χ^2 method [23], features are individually evaluated according to their χ^2 statistic with respect to the classes. For a numeric attribute, the method first requires its range to be discretized into several intervals. The χ^2 value of an attribute (assuming interval and class independence) is defined as:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{c} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{14}$$

where m is the number of intervals or nominal values, c the number of classes, A_{ij} the number of samples in the ith interval, jth class, R_i the number of objects in the ith interval, C_j the number of objects in the jth class, N the total number of objects, and E_{ij} the expected frequency of A_{ij} ($E_{ij} = R_i * C_j/N$). The larger the χ^2 value, the more important the feature.

2.3.4 Symmetrical uncertainty

The Symmetrical Uncertainty [28] (SU) criterion compensates for the inherent bias of IG by dividing it by the sum of the entropies of X and Y:

$$SU(X, Y) = 2 \times \frac{Entropy(X) - Entropy(X|Y)}{Entropy(X) + Entropy(Y)} \tag{15}$$

Due to the correction factor, SU takes values which are normalised to the range [0, 1]. A value of 0 indicates that X and Y are uncorrelated, and 1 means

that the knowledge of one attribute completely predicts the other. Similarly to GR, SU is biased toward features with fewer values.

2.3.5 Relief-F

Relief [18] evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same class (termed *near hit*, H) and different class (termed *near miss*, M). Relief-F extends this idea to dealing with multi-class problems as well as handling noisy and incomplete data. Instead of finding one near miss M, the Relief-F algorithm finds the near miss for each different class $M(C)$ and averages their contribution for updating feature weights as follows:

$$W[A] = W[A] - \frac{d(A, R, H)}{m} + \sum_{C \neq class(R)} \frac{P(C) \times d(A, R, M(C))}{m} \tag{16}$$

where $d(A, R, X)$ is the distance between instance R and instance X for attribute A, $P(C)$ is the prior probability of class C, and m is a normalizing term corresponding to the total number of iterations of the algorithm.

2.3.6 OneR

The OneR classifier [12] learns a one-level decision tree, i.e. it generates a set of rules that test one particular attribute. One branch is assigned for every value of a feature; each branch is assigned the most frequent class. The error rate is then defined as the proportion of instances that do not belong to the majority class of their corresponding branch.

Although OneR is used for classification, it can also be used for evaluating individual features. Features with higher classification rates are considered to

be more significant than those resulting in lower accuracies, and so individual features can be ranked based on this.

3 Estimation of likelihood ratio

Commonly used significance tests, like the Student-t test for univariate data, and Hotelling's T^2 [19, 20], take into account only information about within-source variation and the similarity of the compared items. Thus, the tests provide information on the similarity of items on the basis of their physico-chemical properties. From the forensic point of view, this is inadequate. What is required is information concerning the value of the evidence of these measurements with relation to the proposition that the two samples of glass fragments did, or did not, come from the same source. This requires knowledge about the sources of variability and the rarity of the measured physico-chemical properties in the relevant population. For instance, one would expect refractive index (RI) values from different locations on the same glass object to be very similar. However, equally similar RI values could well be observed from different glass items. Without a wider context it is not possible to ascribe meaning to the observed similarity. Therefore inferences about the source of glass fragments made purely on the basis of similarity of measurements are incomplete. This section outlines several techniques for the estimation of likelihood ratios that attempt to account for the variability and rarity issues through two-level models. Further details are given in [3].

3.1 Exponential model

Consider a two-level random effects model for a random variable X such that $(X_{ij} \mid \mu, \sigma^2) \sim N(\mu, \sigma^2)$ and $(\mu \mid \alpha) \sim Exp(\alpha)$ with probability density function

$$f(\mu \mid \alpha) = \alpha \exp(-\alpha\mu).$$

Let $\{x_{ij}, i = 1, \ldots, m,\ j = 1, \ldots, k\}$ be a random sample from this model of k observations from each of m groups. In the glass fragment analysis domain, a group corresponds to a set of k measurements from a single source, such as from a single window. Denote the m group means by $\bar{x}_1, \ldots, \bar{x}_m$ where $\bar{x}_i = \sum_{j=1}^{k} x_{ij}/k$. The overall mean is denoted $\bar{x}_{\cdot}$ with $\bar{x}_{\cdot} = \sum_{i=1}^{m} \sum_{j=1}^{k} x_{ij}/km$.

Data $\mathbf{y_1} = \{y_{1j}, j = 1, \ldots, n_c\}$ of n_c observations from one group from a crime scene (control data) and data $\mathbf{y_2} = \{y_{2j}, j = 1, \ldots, n_s\}$ of n_s observations from a group associated with a suspect (recovered data) are obtained. The value, V, of the evidence of these data is to be determined.

The value of evidence E in comparing the probabilities of the truth of two propositions, H_p for the prosecution and H_d for the defence say, is taken to be the factor which converts the odds in favour of H_p, relative to H_d, prior to consideration of E, to the odds in favour of H_p, relative to H_d, posterior to consideration of E. From the odds form of Bayes' Theorem, the value of the evidence can be seen to be the likelihood ratio

$$V = \frac{Pr(E \mid H_p)}{Pr(E \mid H_d)}$$

The exponential distribution is investigated as it is not easy to transform

to a normal distribution and because a theoretical value for the likelihood ratio may be obtained against which various estimative procedures may be compared.

Some intermediate calculations and notation are required.

Let $\bar{y}_1 = \sum_{j=1}^{n_c} y_{1j}/n_c$ and $\bar{y}_2 = \sum_{j=1}^{n_s} y_{2j}/n_s$ denote the means of the crime and suspect data, respectively. Let $s_{y1}^2 = \sum_{j=1}^{n_c}(y_{1j} - \bar{y}_1)^2/(n_c - 1)$ and $s_{y2}^2 = \sum_{j=1}^{n_s}(y_{2j} - \bar{y}_2)^2/(n_s - 1)$ denote the variances of the crime and suspect data, respectively.

The within-group variance σ^2 of the underlying population is assumed known. Its value is taken to be $s_w^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}(x_{ij} - \bar{x}_i)^2/(mk - m)$. The between-group variance of the underlying population is also assumed known. Its value is taken to be $s_b^2 = \sum_{i=1}^{m}(\bar{x}_i - \bar{x}_.)^2/(m-1) - s_w^2/k$.

The expectation of μ, an exponentially distributed random variable, is $1/\alpha$. The parameter α is estimated by $(\bar{x})^{-1}$. The variance of μ is $1/\alpha^2$.

The value of the evidence $(\mathbf{y_1}, \mathbf{y_2})$ is given by

$$\begin{aligned} V &= \frac{\int f(\mathbf{y_1}, \mathbf{y_2} \mid \mu, \sigma^2) f(\mu \mid \alpha) d\mu}{\int f(\mathbf{y_1} \mid \mu, \sigma^2) f(\mu \mid \alpha) d\mu \int f(\mathbf{y_2} \mid \mu, \sigma^2) f(\mu \mid \alpha) d\mu} \\ &= \frac{\int f(\mathbf{y_1} \mid \mu, \sigma^2) f(\mathbf{y_2} \mid \mu, \sigma^2) f(\mu \mid \alpha) d\mu}{\int f(\mathbf{y_1} \mid \mu, \sigma^2) f(\mu \mid \alpha) d\mu \int f(\mathbf{y_2} \mid \mu, \sigma^2) f(\mu \mid \alpha) d\mu}. \end{aligned}$$

This is the likelihood ratio, where H_p is assumed to be true for the numerator (i.e. the suspect was at the crime scene, so $\mathbf{y_1}, \mathbf{y_2}$ are from the same source) and H_d is assumed to be true for the denominator (i.e. the two samples are assumed to be from different sources, so $\mathbf{y_1}, \mathbf{y_2}$ are independent). When $\mathbf{y_1}$ and $\mathbf{y_2}$ come from the same source, as is assumed in the numerator, they are

dependent within the marginal distribution. Following the argument in [22], $\mathbf{y_1}$ and $\mathbf{y_2}$ can be transformed to independent statistics $(\bar{y}_1 ... \bar{y}_2, w)$, with unit Jacobian. The numerator becomes $f(\bar{y}_1 - \bar{y}_2) \int f(w \mid \mu) f(\mu \mid \alpha) d\mu$. The value V of the evidence is then (after simplification):

$$V = \frac{1}{\alpha\ \sigma_{12}\ \sqrt{2\pi}} \exp\Big[-\frac{1}{2\sigma_{12}^2}\big(\bar{y}_1 - \bar{y}_2\big)^2 + \frac{\alpha}{2}\Big\{2(\bar{y}_1 + \bar{y}_2 - w) + \alpha\sigma_3^2 - \alpha\sigma^2\big(\frac{1}{n_c} + \frac{1}{n_s}\big)\Big\}\Big]. \tag{17}$$

where

$$\begin{aligned} \sigma_{12}^2 &= \sigma^2\Big(\frac{1}{n_c} + \frac{1}{n_s}\Big); \\ \sigma_3^2 &= \frac{\sigma^2}{n_c + n_s} + \frac{1}{\alpha^2}; \\ w &= (n_c\bar{y}_1 + n_s\bar{y}_2)/(n_c + n_s). \end{aligned}$$

Further information regarding the derivation of V can be found in [3]. For estimations, the parameters α and σ^2 are replaced by their estimates from the population data $\{x_{ij}, i = 1, \ldots, m; j = 1, \ldots, k\}$, namely, $(\bar{x})^{-1}$ and s_w^2, respectively. If the between-group distribution is assumed to be exponential then an estimate of the value of evidence in a particular case with crime data $\mathbf{y}_1$ and suspect data $\mathbf{y}_2$ may be obtained with substitution of the appropriate numerical values for $\bar{y}_1$ and $\bar{y}_2$ in (17).

3.2 Biweight kernel estimation

The use of a kernel density estimate based on the normal distribution is difficult when there is an achievable lower bound to the range of the variable being modelled and the data are highly positively skewed so that much of the data

are close to the lower bound. In the example to be discussed here, the lower bound is zero and a kernel based on a normal distribution is very inaccurate close to this lower bound. A more appropriate approach for modelling a highly positively skewed distribution is the use of a biweight kernel [33] with a boundary kernel for use when the kernel comes close to the lower bound of the range of the random variable, in this case zero. The biweight kernel $K(z)$ is defined as

$$K(z) = \frac{15}{16}(1 - z^2)^2; \;\; \mid z \mid < 1. \tag{18}$$

This kernel is used to model the between-group distribution using the sample means $\{\bar{x}_1, \ldots, \bar{x}_m\}$. A general biweight kernel, with smoothing parameter h, and with a between-group variance of τ^2 is given by

$$\frac{1}{h\tau} K\Big(\frac{\mu - \bar{x}}{h\tau}\Big) = \frac{15}{16h\tau}\Big\{1 - \Big(\frac{\mu - \bar{x}}{h\tau}\Big)^2\Big\}^2; \;\; \bar{x} - h\tau < \mu < \bar{x} + h\tau. \tag{19}$$

There are two candidates for the estimation of the between-group variance,

(i) $s_b^2 = \sum_{i=1}^{m} (\bar{x}_i - \bar{x}_.)^2/(m-1) - s_w^2/k$,

(ii) $1/(\bar{x})^2$,

the least-squares estimate and the method of moments estimate, respectively, of τ^2, the between-group variance.

The problem of a fixed lower bound at zero is tackled with a boundary kernel. When an observation, $\bar{x}$, is close to zero, a different kernel, known as the boundary kernel [33], is used. Closeness is defined as $\bar{x} < h\tau$. For $\bar{x} > h\tau$, the biweight kernel (19) is used. For $\bar{x} < h\tau$, a boundary kernel

$$K_h(z) = \frac{\nu_2 - \nu_1 z}{\nu_0 \nu_2 - \nu_1^2} K(z) \tag{20}$$

is used where $K(z)$ is as given in (18). For ease of notation, denote $h\tau$ by δ. The terms ν_0, ν_1 and ν_2 are constants, functions of δ. For the kernel (18) these are defined as

$$\nu_t = \int_{-1}^{\delta} z^t K(z) dz, \quad t = 0, 1, 2,$$

where the dependency in ν on δ is suppressed. They can be shown to be

$$\begin{aligned}
\nu_2 &= \frac{1}{14}\{1 + \frac{1}{8}\delta^3(35 - 42\delta^2 + 15\delta^4)\}, \\
\nu_1 &= \frac{5}{32}\{\delta^2(3 - 3\delta^2 + \delta^4) - 1\}, \\
\nu_0 &= \frac{1}{2} + \frac{15}{16}(\delta - \frac{2}{3}\delta^3 + \frac{1}{5}\delta^5).
\end{aligned}$$

In practice, the factor $(\nu_2 - \nu_1 z)/(\nu_0\nu_2 - \nu_1^2)$ is close to 1.

An optimal value of the smoothing parameter h is given by

$$h_{opt} = \Big(\frac{1}{7}\Big)^{-\frac{2}{5}} \Big(\frac{15}{21}\Big)^{\frac{1}{5}} \Big\{ \int f''(x)^2 dx \Big\}^{-\frac{1}{5}} m^{-\frac{1}{5}}$$

[31]. Then, it can be shown that, when $f(x) = \alpha \exp\{-\alpha x\}$,

$$h_{opt} = \Big(\frac{70}{m}\Big)^{\frac{1}{5}} \alpha^{-1}$$

which can be estimated by

$$h_{opt} = \Big(\frac{70}{m}\Big)^{\frac{1}{5}} \bar{x}.$$

3.3 Likelihood ratio with biweight and boundary kernels

Recall that the likelihood ratio is defined as

$$\begin{aligned}
V &= \frac{Pr(E \mid H_p)}{Pr(E \mid H_d)} \\
&= \frac{f(\bar{y}_1 - \bar{y}_2) \int f(w \mid \mu) f(\mu \mid \alpha) d\mu}{\int f(\mathbf{y_1} \mid \mu, \sigma^2) f(\mu \mid \alpha) d\mu \int f(\mathbf{y_2} \mid \mu, \sigma^2) f(\mu \mid \alpha) d\mu}
\end{aligned}$$

First, consider the denominator and the factor which is associated with the crime sample $\{y_{1i}, i = 1, \ldots, n_c\}$. Denote this as D_c. This may be written as

$$D_c = \int f(y_{11}, \ldots, y_{1n_c} \mid \mu,\ \sigma^2) f(\mu \mid \alpha) d\mu.$$

The factor associated with the suspect sample may be derived analogously and denote this as D_s. The first term, D_c, in the denominator, with the biweight kernel (19) used for $f(\mu \mid \alpha)$, is given by

$$\begin{aligned} D_c &= \int f(\bar{y}_1 \mid \mu,\ \sigma^2) f(\mu \mid \alpha) d\mu \\ &= \frac{\sqrt{n_c}}{\sigma\sqrt{2\pi}} \int \exp\Big\{ -\frac{n_c}{2\sigma^2}(\bar{y}_1 - \mu)^2 \Big\} \Big[\frac{15}{16\ m\ h\ \tau} \sum_{i=1}^{m} \Big\{ 1 - \Big(\frac{\mu - \bar{x}_i}{h\ \tau} \Big)^2 \Big\}^2 \Big] d\mu \\ &= \frac{15\ \sqrt{n_c}}{16\ m\ \sigma\sqrt{2\pi}} \sum_{i=1}^{m} \int_{-1}^{1} (1 - z_i^2)^2 \exp\Big\{ -\frac{n_c}{2\sigma^2} (\bar{y}_1 - (\bar{x}_i + h\ \tau\ z_i))^2 \Big\} dz_i. \end{aligned}$$

Similarly, the second term, D_s, in the denominator, with the biweight kernel (19) used for $f(\mu \mid \alpha)$, is given by

$$\begin{aligned} D_s &= \int f(\bar{y}_2 \mid \mu,\ \sigma^2) f(\mu \mid \alpha) d\mu \\ &= \frac{15\ \sqrt{n_s}}{16\ m\ \sigma\sqrt{2\pi}} \sum_{i=1}^{m} \int_{-1}^{1} (1 - z_i^2)^2 \exp\Big\{ -\frac{n_s}{2\sigma^2} (\bar{y}_2 - (\bar{x}_i + h\ \tau\ z_i))^2 \Big\} dz_i. \end{aligned}$$

The numerator, N, is given by

$$\begin{aligned} N &= f(\bar{y}_1 - \bar{y}_2) \int f(w \mid \mu) f(\mu \mid \alpha) d\mu \\ &= \frac{1}{\sigma_{12}\ \sqrt{2\pi}} \exp\Big\{ -\frac{1}{2\sigma_{12}^2} (\bar{y}_1 - \bar{y}_2)^2 \Big\} \frac{15}{16\ m\ \sigma_3\ \sqrt{2\pi}} \\ &\quad \sum_{i=1}^{m} \int_{-1}^{1} (1 - z_i^2)^2 \exp\Big[-\frac{1}{2\sigma_3^2} \big\{ w - (\bar{x}_i + h\ \tau\ z_i) \big\}^2 \Big] dz_i. \end{aligned}$$

It can be shown that the likelihood ratio is given by the ratio of N to the product of D_c and D_s. Numerical evaluation of the likelihood ratio may then

be made with the substitution of σ by s_w, τ by s_b and h by its optimal value $(70/m)^{1/5}\bar{x}$.

There is a boundary effect when an $(\bar{x}_i, i = 1, \ldots, m)$ is within $h\tau$ of zero. For these $\bar{x}_i$, the kernel expression

$$\Big\{1 - \Big(\frac{\mu - \bar{x}_i}{h\tau}\Big)^2\Big\}^2 = \Big\{1 - z_i^2\Big\}^2$$

has to be adjusted with the factor $(\nu_2 - \nu_1 z)/(\nu_0\nu_2 - \nu_1^2)$, where $z_i = (\mu - \bar{x}_i)/(h\tau)$ and ν_0, ν_1, ν_2 are as in (20), to give

$$\frac{(\nu_2 - \nu_1 z_i)}{\nu_0\nu_2 - \nu_1^2)}\Big\{1 - z_i^2\Big\}^2$$

which can be written as

$$(a - bz_i)\Big\{1 - z_i^2\Big\}^2$$

where $a = \nu_2/(\nu_0\nu_2 - \nu_1^2)$ and $b = \nu_1/(\nu_0\nu_2 - \nu_1^2)$. Define an indicator function $\gamma(z_i)$ such that

$$\begin{aligned} \gamma(z_i) &= 1 \text{ if } x_i \geq h\tau, \\ &= (a - bz_i) \text{ if } x_i < h\tau. \end{aligned}$$

Then the likelihood ratio $N/(D_c D_s)$ can be adapted to account for boundary effects to give a value for the evidence of

$$\frac{1}{\sigma_{12}\sqrt{2\pi}} \exp\Big\{ - \frac{1}{2\sigma_{12}^2}(\bar{y}_1 - \bar{y}_2)^2\Big\} \frac{15}{16\, m\, \sigma_3\, \sqrt{2\pi}} \times$$
$$\sum_{i=1}^{m} \int_{-1}^{1} \gamma(z_i)(1 - z_i^2)^2 \exp\Big[- \frac{1}{2\sigma_3^2}\big\{w - (\bar{x}_i + h\,\tau\, z_i)\big\}^2\Big] dz_i.$$

divided by the product of

$$\frac{15\ \sqrt{n_c}}{16\ m\ \sigma\sqrt{2\pi}} \sum_{i=1}^{m} \int_{-1}^{1} \gamma(z_i)(1-z_i^2)^2 \exp\Big\{ -\frac{n_c}{2\sigma^2}(\bar{y}_1 - (\bar{x}_i + h\ \tau\ z_i))^2 \Big\} dz_i$$

and

$$\frac{15\ \sqrt{n_s}}{16\ m\ \sigma\sqrt{2\pi}} \sum_{i=1}^{m} \int_{-1}^{1} \gamma(z_i)(1-z_i^2)^2 \exp\Big\{ -\frac{n_s}{2\sigma^2}(\bar{y}_2 - (\bar{x}_i + h\ \tau\ z_i))^2 \Big\} dz_i.$$

3.4 Adaptive kernel

The value of the evidence, when the between-group distribution is taken to be non-normal and is estimated by a normal kernel function as described in [1], equation (10.12), is adapted to allow for the correlation between the control and recovered data $\bar{y}_1$ and $\bar{y}_2$ if they are assumed, as in the numerator, to come from the same source. This expression is then extended to an adaptive kernel, where the smoothing parameter is dependent on x_i and is thus denoted h_i.

The numerator for equation (10.12) is

$$\frac{1}{m}(2\pi)^{-1}\{(\frac{n_c+n_s}{n_c\ n_s})\sigma^2\}^{-1/2}\{\tau^2 + \frac{\sigma^2}{n_c+n_s}\}^{-1/2}(h_i^2\tau^2)^{-1/2}$$
$$\{(\tau^2 + \frac{\sigma^2}{n_c+n_s})^{-1} + (h_i^2\tau^2)^{-1}\}^{-1/2} \exp\{-\frac{1}{2}(\bar{y}_1 - \bar{y}_2)^2[(\frac{n_c+n_s}{n_c\ n_s})\sigma^2]^{-1}\}$$
$$\sum_{i=1}^{m} \exp\{-\frac{1}{2}(w - \bar{x}_i)^2(\tau^2 + \frac{\sigma^2}{n_c+n_s} + h_i^2\tau^2)^{-1}\}.$$

The first term in the denominator of equation (10.12) is

$$\frac{1}{m}(2\pi)^{-1/2}\{\tau^2 + \frac{\sigma^2}{n_c}\}^{-1/2}(h_i 2\ \tau^2)^{-1/2}$$

$$\{(\tau^2 + \frac{\sigma^2}{n_c})^{-1} + (h_i^2\tau^2)^{-1}\}^{-1/2}$$

$$\sum_{i=1}^{m} \exp\{-\frac{1}{2}(\bar{y}_1 - \bar{x}_i)^2(\tau^2 + \frac{\sigma^2}{n_c} + h_i^2\tau^2)^{-1}\}.$$

The second term in the denominator of equation (10.12) is

$$\frac{1}{m}(2\pi)^{-1/2}\{\tau^2 + \frac{\sigma^2}{n_s}\}^{-1/2}(h_i^2\ \tau^2)^{-1/2}$$

$$\{(\tau^2 + \frac{\sigma^2}{n_s})^{-1} + (h_i^2\tau^2)^{-1}\}^{-1/2}$$

$$\sum_{i=1}^{m} \exp\{-\frac{1}{2}(\bar{y}_2 - \bar{x}_i)^2(\tau^2 + \frac{\sigma^2}{n_s} + h_i^2\tau^2)^{-1}\}.$$

The constant term in the ratio is then:

$$\frac{m\Big\{n_c\tau^2(h_i^2+1)+\sigma^2\Big\}^{1/2}\Big\{n_s\tau^2(h_i^2+1)+\sigma^2\Big\}^{1/2}}{\sigma\Big\{(n_c+n_s)\tau^2(h_i^2+1)+\sigma^2\Big\}^{1/2}}.$$

The remaining term, that involving $\bar{y}_1, \bar{y}_2$ and $\bar{x}_i$, is the ratio of

$$\exp\{-\frac{1}{2}(\bar{y}_1-\bar{y}_2)^2(\sigma^2(\frac{1}{n_c}+\frac{1}{n_s}))^{-1}\}\sum_{i=1}^{m}\exp\{-\frac{1}{2}(w-\bar{x}_i)^2(\tau^2+\frac{\sigma^2}{n_c+n_s}+h_i^2\tau^2)^{-1}\}$$

to

$$\sum_{i=1}^{m}\exp\{-\frac{1}{2}(\bar{y}_1-\bar{x}_i)^2(\tau^2+\frac{\sigma^2}{n_c}+h_i^2\tau^2)^{-1}\}\sum_{i=1}^{m}\exp\{-\frac{1}{2}(\bar{y}_2-\bar{x}_i)^2(\tau^2+\frac{\sigma^2}{n_s}+h_i^2\tau^2)^{-1}\}.$$

3.4.1 Adaptive smoothing parameter

The adaptive smoothing parameter h_i is estimated using the procedure outlined in [31].

First, find a pilot estimate $\tilde{f}(x)$ that satisfies $\tilde{f}(x_i) > 0$ for all i. This is achieved by standard kernel density estimation with Gaussian kernels [31]. Then define the smoothing parameter h_i by

$$h_i = \left\{\tilde{f}(x_i)/g\right\}^{-\beta}$$

where g is the geometric mean of the $\tilde{f}(x_i)$:

$$\log g = m^{-1} \sum \log \tilde{f}(x_i)$$

and β is a sensitivity parameter, a number satisfying $0 \leq \beta \leq 1$.

4 Application

In order to evaluate the feature selection methods and LR estimators, a forensic dataset was obtained from the Forensic Research Institute, Krakow, Poland. An overview of the overall system can be found in Fig. 1.

One large piece of glass from each of 200 glass objects from various sources (including float and container glass) was selected. Each of these 200 pieces was wrapped in a sheet of grey paper and further fragmented. The fragments from each piece were placed in a plastic Petri dish. Four glass fragments, of linear dimension less than 0.5mm with surfaces as smooth and flat as possible, were selected for examination with the use of an SMXX Carl Zeiss (Jena, Germany) optical microscope (magnification 100×).

4.1 Fragment elemental analysis

The four selected glass fragments were placed on self-adhesive carbon tabs on an aluminium stub and then carbon coated using an SCD sputter (Bal-Tech, Switzerland). The prepared stub was mounted in the sample chamber of a scanning electron microscope. Analysis of the elemental content of each glass fragment was carried out using a scanning electron microscope (JSM-5800 Jeol, Japan), with an energy dispersive X-ray spectrometer (Link ISIS 300, Oxford Instruments Ltd., United Kingdom).

Three replicate measurements were taken from different areas on each of the four fragments, making twelve measurements from each glass object, but only four independent measurements. The four means of the measurements were used for the analysis. The measurement conditions were accelerating voltages 20kV , life time $50s$, magnification 1000 - 2000×, and the calibration element was cobalt. The SEMQuant option (part of the software LINK ISIS, Oxford Instruments Ltd, United Kingdom) was used in the process of determining the percentage of particular elements in a fragment. The option applied a ZAF correction procedure, which takes into account corrections for the effects of difference in the atomic number (Z), absorption (A) and X-ray fluorescence (F).

The selected analytical conditions allowed the determination of all elements having an atomic number greater than 5 (Boron) when they were present at levels greater than the detection limits. However, only the concentrations of oxygen (O), sodium (Na), magnesium (Mg), aluminium (Al), silicon (Si), potassium (K), calcium (Ca) and iron (Fe) are considered further in this paper as

glass is essentially a silicon oxide with sodium and/or calcium added to create a commonly produced glass, and potassium, magnesium, aluminium and iron added to stabilise its structure and modify its physico-chemical properties. Histograms of the distributions of the data can be found in Fig. 2.

4.2 Data preparation

As two of the feature selection algorithms require fuzzy sets to be defined for each element in the dataset in order to maximise the use of information contained in the real-valued variables, further processing is required. Note that the attributes still take real values, and hence no discretization is performed. For this set of experiments, five fuzzy sets per feature were derived automatically based on the mean and standard deviation as seen in Fig. 3. The value λ was set at 0.7. There is no theoretical argument as to why five sets should be chosen, although psychological research suggests that 5, 7 or 9 categories should be used, mimicking human cognition [26]. To minimize computational effort, only 5 sets are defined for this application.

4.3 Feature selection

The feature selection methods outlined previously were applied to the processed glass data in order to select a single attribute for use in the univariate LR estimators. The dataset containing the full set of elements was processed by each method, resulting in a ranking of these features. The top ranked feature/element for each method was then selected, and the data reduced to this feature only.

4.4 Estimators

The performance of four procedures for estimating the likelihood ratio is compared. The procedures estimate the between-group distributions with a normal distribution, an exponential distribution, a normal adaptive kernel and a biweight kernel with a boundary condition.

5 Experimentation

This section presents the results of a comparison of both different feature selection methods and different likelihood ratio estimation procedures, in order to gauge the utility of both. In the experimentation, two situations need to be considered; namely, when the control and recovered data are from the same source and when they are from different sources. For same-source comparisons, the control and recovered data are taken from the same group by splitting the group into two equally-sized, non-overlapping halves (containing two measurements each). For different-source comparisons, the control and recovered data are entire groups selected from different sources.

5.1 Feature selection

Table 1 (summarized in table 2) presents the ordering of features as determined by several leading measures of feature significance: fuzzy-rough feature selection (FRFS), fuzzy entropy (FuzEnt), χ^2, gain ratio (GR), information gain (IG), OneR, Relief-F and symmetrical uncertainty (SU). It can be seen that FRFS, IG and OneR select aluminium; FuzEnt and GR select sodium; χ^2 and SU select potassium; and Relief-F selects magnesium. Based on these selections

and corresponding data reductions, the four estimators are applied.

5.2 Likelihood ratio estimation

There are 200 within-group comparisons of control and recovered data and $200 \times 199/2 = 19,900$ between-group comparisons For the 200 within-group comparisons, the likelihood ratio should be greater than 1 and for the 19,900 between-group comparisons, the likelihood ratio should be less than 1. Results are recorded for a normal kernel estimation (nn), an exponential kernel estimation (exp), an adaptive kernel estimation for $\beta = 0, 0.1, 0.2$ and 0.5 and a biweight kernel estimation (b).

The results for the within-group and between-group estimations are shown in tables 3 and 4 respectively. It can be seen that, overall, the results produced using the aluminium data are superior than those generated using the other considered elements. This is the element commonly chosen by domain experts for univariate modelling [3]. The within-group estimation of likelihood ratio is as good as or better than the others, and the between-group estimation is more accurate. The results also indicate that magnesium is a very informative feature for likelihood ratio estimation.

The level of false positives (important from a forensic viewpoint) is rather high. This could be the result of several factors. The SEM-EDX method delivers information about the main elemental content of glass, thus differences between glass objects are very small. Moreover, soda-lime-silica glass (the most common) has a strict composition and production recipes, used in many factories, are very similar. Also, the data itself is created mostly from car windows

and building windows, which have extremely similar elemental concentrations. Sodium results are relatively bad as the concentration of Na is restricted by the definition of Na-Ca-Si glass. Potassium results are also unsatisfactory because it is only present in glass objects that have special optical properties, and so is absent or undetectable in many objects (this may explain the poor performance of the exponential model for this element). Aluminium and Magnesium results are relatively superior as these elements are added to glass to create desired features of glass objects and thus their content varies with glass properties.

For this dataset, the feature ranking methods FRFS, IG and OneR have selected the best feature for likelihood estimation. This is in keeping with other investigations that have shown the utility of FRFS, in particular, for this purpose [16]. Although Relief-F selected an alternative feature, the results achieved were almost as accurate. The results show that the normal kernel estimation procedure is inadequate at modelling the underlying data distributions in general. Excepting the case of sodium, it can be seen that even an exponential distribution is a better model as the resulting likelihood ratios are more accurate. The eight methods of feature selection produced four different sets of results as to the top-ranked feature. This illustrates the reason why an expert may be needed to interpret the results. For example, Na, Si, and O are considered by almost any expert to be the least useful elements for discriminating among glass sources. As expected, Si and O do indeed rank as the worst elements. However, the observations for Na are perhaps surprising, which may be as a result of Na ion migration (a well-known problem with SEM-EDX measurements).

The point of the analysis is not to say whether some fragment is likely to have been derived from a glass object or not, but to make statements about to what extent the observations (in this case, of elemental concentrations) lend support to the proposition that the fragment came from the glass object in question, or to the proposition that the fragment came from some other glass object from the relevant population of objects. The likelihood ratio does not provide a direct attribution of source - it provides a measure of the strength of the evidence in support of one or other of the propositions in a case.

Small likelihood ratios would be regarded as very strong support that the particular fragment from which the observations were made had come from some source other than the broken glass object from the crime scene. However, this value, whilst providing very strong support, may not be persuasive enough, one way or the other, to affect the outcome of the case as a whole. The outcome is dependent upon other evidence and court requirements.

6 Conclusion

As a type of evidence, glass can be very useful contact trace material in a wide range of offences including burglaries and robberies, murders, assaults, criminal damage and thefts of and from motor vehicles. In all of these situations, there is the potential for glass fragment transfer, providing a link between suspect and crime. Hence, the correct interpretation of glass evidence is critical in establishing such connections.

Previous work has been based on the use of a normal kernel estimation

procedure for evidence evaluation. However, this may be inadequate when the data is positively skewed and an exponential distribution is thought to be a better model [3]. Three techniques that attempt to alleviate this problem were investigated and found to provide better likelihood ratio estimations.

In this paper, the role of feature selection as an aid to glass analysis was investigated. As the two-level models used were univariate, the task of the selection methods was to determine the most informative feature for use in the models themselves. The results have shown that automated feature selection techniques can indeed aid the choice of variable for further modelling. This choice is a critical factor in the resulting quality of evidence evaluation. Situations are often encountered where many competing variables co-exist. The manual selection of which variable to use may result in subsequent analysis being too subjective. Through the use of feature selection methods, this important decision can be made without expert assistance.

Further work in the area of forensic glass analysis includes fragment identification. The classification of glass samples into their product/use type is important, e.g. in corroborating or disproving an alibi or product tampering. It is expected that not all features involved (in this case, chemicals) will be useful for determining glass type, and hence feature selection methods could be applied to remove this redundant, and possibly misleading, information. In [2], log ratios of oxygen with the seven other variables are used as a standardization procedure and as a transformation to normality. A similar approach could be adopted for the work presented in this paper.

Acknowledgements

This work is funded by the UK EPSRC grant GR/S98603/01. The authors are very grateful to Professor C. G. G. Aitken of the University of Edinburgh, Scotland, UK for his support in this work, while taking full responsibility for the views expressed in this paper. They are also grateful to Dr G. Zadora of the Institute of Forensic Research, Krakow, Poland for the provision of the glass data.

References

[1] C.G.G. Aitken and F. Taroni, *Statistics and the evaluation of evidence for forensic scientists*, 2nd edition, John Wiley and Sons, Ltd, Chichester, 2004.

[2] C.G.G. Aitken, G. Zadora and D. Lucy, A two-level model for evidence evaluation, *Journal of Forensic Sciences*, vol. 52, pp. 412–419, 2007.

[3] C.G.G. Aitken, Q. Shen, R. Jensen and B. Hayes, The evaluation of evidence for exponentially distributed data, *Computational Statistics and Data Analysis*, vol. 51, no. 12, pp. 5682–5693, 2007.

[4] J.G. Bazan, H.S. Nguyen and M.S. Szczuka, A View on Rough Set Concept Approximations, *Fundamenta Informaticae*, 59(2–3), pp. 107–118, 2004.

[5] A. Chouchoulas and Q. Shen, Rough set-aided keyword reduction for text categorisation, *Applied Artificial Intelligence*, vol. 15, no. 9, pp. 843–873, 2001.

[6] M. Dash and H. Liu, Feature Selection for Classification, *Intelligent Data Analysis*, Vol. 1, No. 3, pp. 131–156, 1997.

[7] P. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice Hall, 1982.

[8] D. Dubois and H. Prade, Putting rough sets and fuzzy sets together, in [32] pp. 203–232, 1992.

[9] B. Flury and H. Riedwyl, *Multivariate Statistics: A Practical Approach*, Prentice Hall, 1988.

[10] I. Guyon, S. Gunn, M. Nikravesh and L.A. Zadeh (eds), Feature Extraction: Foundations and Applications, *Studies in Fuzziness and Soft Computing*, Springer-Verlag, New York Inc., 2006.

[11] U. Höhle, Quotients with respect to similarity relations, *Fuzzy Sets and Systems*, 27:31-44, 1988.

[12] R.C. Holte, Very simple classification rules perform well on most commonly used datasets, *Machine Learning*, Vol. 11, No. 1, pp. 63–90, 1993.

[13] E. Hunt, J. Martin and P. Stone, Experiments in Induction, New York Academic Press, 1966.

[14] C.Z. Janikow, Fuzzy Decision Trees: Issues and Methods, *IEEE Transactions on Systems, Man and Cybernetics — Part B: Cybernetics*, **28** 1–14, 1998.

[15] R. Jensen and Q. Shen, Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough Based Approaches, *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 12, pp. 1457–1471, 2004.

[16] R. Jensen and Q. Shen, Fuzzy-Rough Sets Assisted Attribute Selection, *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 1, pp. 73–89, 2007.

[17] R. Jensen and Q. Shen, *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches*, Wiley-IEEE Press, 2008.

[18] I. Kononenko, Estimating attributes: Analysis and Extensions of Relief, Proceedings of the European Conference on Machine Learning, pp. 171–182, 1994.

[19] R.D. Koons and J. Buscaglia, The forensic significance of glass composition and refractive index measurements, *J. Forensic Sci.*, 44(3):496–503, 1999.

[20] R.D. Koons and J. Buscaglia, Interpretation of glass composition measurements: the effects of match criteria on discrimination capability, *J. Forensic Sci.*, 47(3):505–512, 2002.

[21] B. Kosko, Fuzzy entropy and conditioning, *Information Sciences* 40(2): 165–174, 1986.

[22] D.V. Lindley, A problem in forensic science. *Biometrika* 64: 207–213, 1977.

[23] H. Liu and R. Setiono, Chi2: feature evaluation and discretization of numeric attributes, in: Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence, pp. 336–391, 1995.

[24] H. Liu, F. Hussain, C.L. Tan and M. Dash, Discretization: An Enabling Technique, *Data Mining and Knowledge Discovery*, 6(4): 393–423, 2002.

[25] H. Liu and H. Motoda (eds), *Computational Methods of Feature Selection*, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2008.

[26] G.A. Miller, The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information, *Psychological Review*, 63:81–97, 1956.

[27] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishing, Dordrecht, 1991.

[28] W. Press, S.A. Teukolsky, W.T. Vellerlin and B.P. Flannery, *Numerical Recipes in C*, Cambridge University Press, 1988.

[29] J.R. Quinlan, *C4.5: Programs for Machine Learning*, The Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[30] Q. Shen and A. Chouchoulas, A fuzzy-rough approach for generating classification rules, *Pattern Recognition*, 35(11):341–354, 2002.

[31] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall, 1986.

[32] R. Slowinski(ed.), *Intelligent Decision Support*, Kluwer Academic Publishers, Dordrecht, 1992.

[33] M. Wand and M. Jones, *Kernel Smoothing*, Chapman & Hall, London, 1995.

[34] L.A. Zadeh, *Fuzzy sets*, *Information and Control*, 8:338–353, 1965.

List of Figures

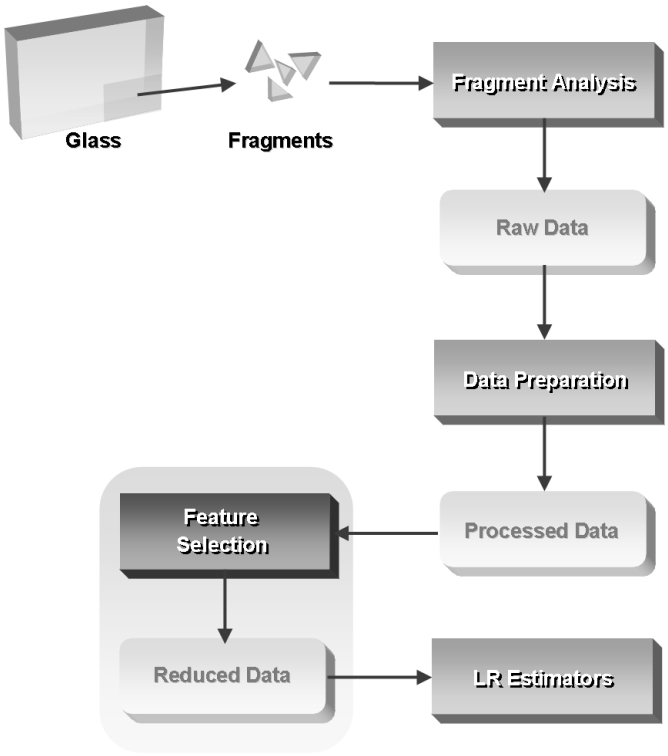

Fig. 1: System overview.

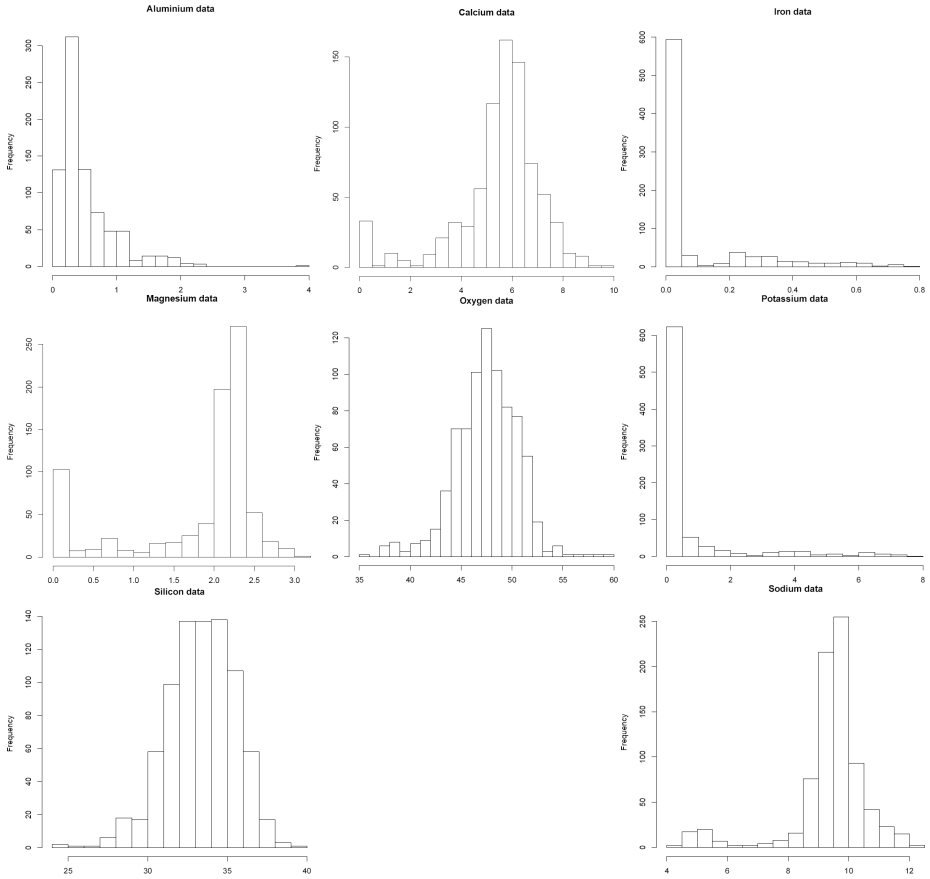

Fig. 2: Data distributions for the eight elements.

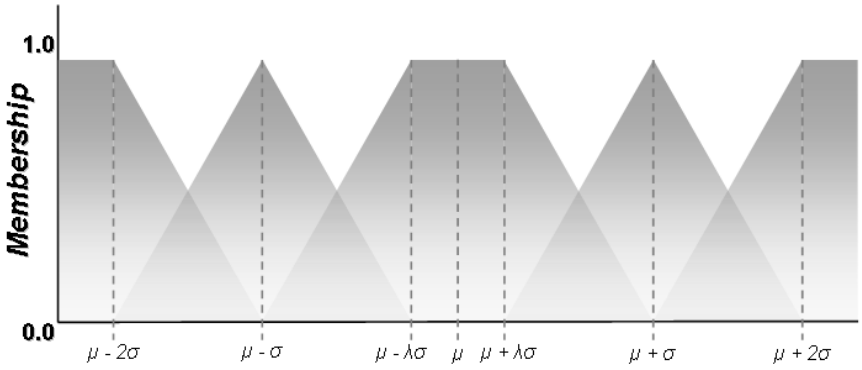

Fig. 3: Fuzzy set construction.

List of Tables

Table 1: Evaluation of features

Element	FRFS	IG	OneR	FuzEnt	GR	Relief-F	χ^2	SU
O	0.003929	0.267	52.50	1.839151	0.211	0.0364	368.4882	0.1615
Na	0.025909	0.597	55.75	1.581143	0.391	0.0894	968.2267	0.3345
Mg	0.025426	0.718	52.13	1.658684	0.367	0.1234	1010.738	0.3589
Al	0.025996	0.843	57.63	1.647498	0.275	0.0661	1167.625	0.3301
Si	0.009267	0.129	41.38	1.801163	0.127	0.0309	150.6826	0.0846
K	0.003118	0.825	55.75	1.582525	0.338	0.0928	1578.554	0.3682
Ca	0.008341	0.511	53.25	1.600595	0.312	0.0881	686.8860	0.2774
Fe	0.022455	0.191	45.50	1.741598	0.145	0.0453	174.0253	0.1136

Table 2: Summary of feature evaluation, where $A > B$ indicates that A is a more informative feature than B, and $A = B$ indicates identical ranking.

Selection method	Feature ranking
FRFS	**Al**>Na>Mg>Fe>Si>Ca>O>K
IG	**Al**>K>Mg>Na>Ca>O>Si>Fe
OneR	**Al**>Na=K>Ca>O>Mg>Fe>Si
FuzEnt	**Na**>K>Ca>Al>Mg>Fe>Si>O
GR	**Na**>Mg>K>Ca>Al>O>Fe>Si
Relief-F	**Mg**>K>Na>Ca>Al>Fe>O>Si
χ^2	**K**>Al>Mg>Na>Ca>O>Fe>Si
SU	**K**>Mg>Na>Al>Ca>O>Fe>Si

Table 3: Summary of likelihood ratios for within-group comparisons.

Likelihood ratio range	nn	exp	β 0.0	0.1	0.2	0.5	b
	Aluminium (Al) (chosen by FRFS, IG and OneR)						
$0-1$	4	4	4	4	4	4	4
$1-10^1$	1	160	184	184	184	185	173
10^1-10^2	183	35	12	12	12	11	22
10^2-10^3	8	1	0	0	0	0	1
10^3-10^4	3	0	0	0	0	0	0
$>10^4$	1	0	0	0	0	0	0
Misclassification	2.0%	2.0%	2.0%	2.0%	2.0%	2.0%	2.0%
	Magnesium (Mg) data (chosen by Relief-F)						
$0-1$	6	6	6	6	6	6	6
$1-10^1$	0	29	157	157	157	157	151
10^1-10^2	165	165	37	37	37	37	43
10^2-10^3	29	0	0	0	0	0	0
10^3-10^4	0	0	0	0	0	0	0
$>10^4$	0	0	0	0	0	0	0
Misclassification	3.0%	3.0%	3.0%	3.0%	3.0%	3.0%	3.0%
	Potassium (K) data (chosen by χ^2 and SU)						
$0-1$	4	43	4	4	4	4	4
$1-10^1$	174	138	181	181	181	181	170
10^1-10^2	13	3	15	15	15	15	26
10^2-10^3	4	9	0	0	0	0	0
10^3-10^4	2	4	0	0	0	0	0
$>10^4$	3	3	0	0	0	0	0
Misclassification	2.0%	21.5%	2.0%	2.0%	2.0%	2.0%	2.0%
	Sodium (Na) data (chosen by FuzEnt and GR)						
$0-1$	5	3	7	7	7	7	4
$1-10^1$	179	6	183	183	183	183	183
10^1-10^2	6	191	10	10	10	10	13
10^2-10^3	5	0	0	0	0	0	0
10^3-10^4	5	0	0	0	0	0	0
$>10^4$	0	0	0	0	0	0	0
Misclassification	2.5%	1.5%	3.5%	3.5%	3.5%	3.5%	2.0%

Table 4: Summary of likelihood ratios for between-group comparisons.

Likelihood ratio range	nn	exp	β				b
			0.0	0.1	0.2	0.5	
			Aluminium (Al) (chosen by FRFS, IG and OneR)				
$0-1$	10661	12548	12669	12662	12652	12518	13924
$1-10^1$	2958	7031	7172	7180	7190	7319	5866
10^1-10^2	6243	320	59	58	58	63	120
10^2-10^3	34	1	0	0	0	0	0
10^3-10^4	4	0	0	0	0	0	0
$>10^4$	0	0	0	0	0	0	0
Misclassification	46.4%	36.9%	36.3%	36.4%	36.4%	37.1%	30.0%
			Magnesium (Mg) data (chosen by Relief-F)				
$0-1$	10955	11220	12020	12016	12003	11965	12408
$1-10^1$	1673	2032	6969	6868	6707	5937	7169
10^1-10^2	6983	6648	911	1016	1190	1998	323
10^2-10^3	289	0	0	0	0	0	0
10^3-10^4	0	0	0	0	0	0	0
$>10^4$	0	0	0	0	0	0	0
Misclassification	44.9%	43.6%	39.6%	39.6%	39.7%	39.9%	37.6%
			Potassium (K) data (chosen by χ^2 and SU)				
$0-1$	6463	7931	6989	6988	6986	6971	7861
$1-10^1$	2722	11881	12869	12870	12872	12887	11960
10^1-10^2	10696	44	40	41	41	41	78
10^2-10^3	8	33	2	1	1	1	1
10^3-10^4	8	9	0	0	0	0	0
$>10^4$	3	2	0	0	0	0	0
Misclassification	67.5%	60.1%	64.9%	64.9%	64.9%	65.0%	60.5%
			Sodium (Na) data (chosen by FuzEnt and GR)				
$0-1$	8540	6543	9414	9414	9410	9345	7609
$1-10^1$	11239	2879	10452	10451	10455	10519	6614
10^1-10^2	83	10478	34	35	35	36	5677
10^2-10^3	18	0	0	0	0	0	0
10^3-10^4	20	0	0	0	0	0	0
$>10^4$	0	0	0	0	0	0	0
Misclassification	57.1%	67.1%	52.7%	52.7%	52.7%	53.0%	61.8%